Probabilistic Bag-Of-Hyperlinks Model for Entity Linking

Octavian-Eugen Ganea
Dept. of Computer Science
ETH Zurich, Switzerland
ganeao@inf.ethz.ch

Marina Ganea[*]
Dept. of Computer Science
ETH Zurich, Switzerland
marinah@google.com

Aurelien Lucchi
Dept. of Computer Science
ETH Zurich, Switzerland
alucchi@inf.ethz.ch

Carsten Eickhoff
Dept. of Computer Science
ETH Zurich, Switzerland
ecarsten@inf.ethz.ch

Thomas Hofmann
Dept. of Computer Science
ETH Zurich, Switzerland
thomaho@inf.ethz.ch

ABSTRACT

Many fundamental problems in natural language processing rely on determining what entities appear in a given text. Commonly referenced as *entity linking*, this step is a fundamental component of many NLP tasks such as text understanding, automatic summarization, semantic search or machine translation. Name ambiguity, word polysemy, context dependencies and a heavy-tailed distribution of entities contribute to the complexity of this problem.

We here propose a probabilistic approach that makes use of an effective graphical model to perform collective *entity disambiguation*. Input mentions (i.e., linkable token spans) are disambiguated jointly across an entire document by combining a document-level prior of entity co-occurrences with local information captured from mentions and their surrounding context. The model is based on simple sufficient statistics extracted from data, thus relying on few parameters to be learned.

Our method does not require extensive feature engineering, nor an expensive training procedure. We use loopy belief propagation to perform approximate inference. The low complexity of our model makes this step sufficiently fast for real-time usage. We demonstrate the accuracy of our approach on a wide range of benchmark datasets, showing that it matches, and in many cases outperforms, existing state-of-the-art methods.

Keywords

Entity linking; Entity disambiguation; Wikification; Probabilistic graphical models; Approximate inference; Loopy belief propagation

1. INTRODUCTION

Digital systems are producing increasing amounts of data every day. With daily global volumes of several terabytes of newly textual content, there is a growing need for automatic methods for text aggregation, summarization, and, eventually, semantic understanding. Entity linking is a key step towards these goals as it reveals the semantics of spans of text that refer to real-world entities. In practice, this is achieved by establishing a mapping between potentially ambiguous surface forms of entities and their canonical representations such as corresponding Wikipedia[1] articles or Freebase[2] entries. Figure 1 illustrates the difficulty of this task when dealing with real-world data. The main challenges arise from word ambiguities inherent to natural language: surface form *synonymy*, i.e., different spans of text referring to the same entity, and *homonymy*, i.e., the same name being shared by multiple entities.

Figure 1: An entity disambiguation problem showcasing five given mentions and their potential entity candidates.

We here describe and evaluate a novel light-weight and fast alternative to heavy machine-learning approaches for document-level entity disambiguation with Wikipedia. Our model is primarily based on simple empirical statistics acquired from a training dataset and relies on a very small number of learned parameters. This has certain advantages like a very fast training procedure that can be applied to massive amounts of data, as well as a better understanding of the model compared to increasingly popular deep learning architectures (e.g., He et al. [14]). As a prerequisite, we

---

[*]Currently at Google Inc.

[1]http://en.wikipedia.org/

[2]https://www.freebase.com/

assume that a given input set of mentions was already discovered via a *mention detection* procedure[3]. Our starting point is the natural assumption that each entity depends (i) on its mention, (ii) its neighboring local contextual words, and (iii) on other entities that appear in the same document.

In order to enforce these conditions, we rely on a conditional probabilistic model that consists of two parts: (1) the likelihood of a candidate entity given the referring token span and its surrounding context, and (2) the prior joint distribution of the candidate entities corresponding to all the mentions in a document. Our model relies on the max-product algorithm to collectively infer entities for all mentions in a given document.

We further illustrate these modeling decisions. In the example depicted in Figure 1, each highlighted mention constrains the set of possible entity candidates to a limited size set, yet leaves a significant level of ambiguity. However, there is one collective way of linking that is jointly consistent with all the chosen entities and supported by contextual cues. Intuitively, the related entities `Thomas_Müller` and `Germany_national_football_team` are likely to appear in the same document, especially in the presence of contextual words related to soccer, like "team" or "goal".

Our main contributions are outlined below: **(1)** We employ rigorous probabilistic semantics for the *entity disambiguation* problem by introducing a principled probabilistic graphical model that requires a simple and fast training procedure. **(2)** At the core of our joint probabilistic model, we derive a minimal set of potential functions that proficiently explain statistics of observed training data. **(3)** Throughout a range of experiments performed on several standard datasets using the Gerbil platform [37], we demonstrate competitive or state of the art quality compared to some of the best existing approaches. **(4)** Moreover, our training procedure is solely based on publicly available Wikipedia hyperlink statistics and the method does not require extensive hyperparameter tuning, nor feature engineering, making this paper a self-contained manual of implementing an entity disambiguation system from scratch.

The remainder of this paper is structured as follows: Section 2 briefly discusses relevant entity linking literature. Section 3 formally introduces our probabilistic graphical model and details the initialization and learning procedure of the model's parameters. Section 4 describes the inference process used for collective entity resolution. Section 5 empirically demonstrates the merits of the proposed method on multiple standard collections of manually annotated documents. Finally, in Section 6, we conclude with a summary of our findings and an overview of ongoing and future work.

2. RELATED WORK

There is a substantial body of existing work dedicated to the task of entity linking with Wikipedia (Wikification). We can identify four major paradigms of how this challenge is approached.

Local models consider the individual context of each entity mention in isolation in order to reduce the size of the decision space. In one of the early entity linking papers, Mihalcea and Csomai [21] propose an entity disambiguation scheme

based on similarity statistics between the mention context and the entity's Wikipedia page. Milne and Witten [22] further refine their scheme with special focus on the mention detection step. Bunescu and Pasca [2] present a Wikipedia-driven approach, making use of manually created resources such as redirect and disambiguation pages. Dredze et al. [7] cast the entity linking task as a retrieval problem, treating mentions and their contexts as queries, and ranking candidate entities according to their likelihood of being referred to.

Global models attempt to jointly disambiguate all mentions in a document based on the assumption that the underlying entities are correlated and consistent with the main topic of the document. While this approach tends to result in superior accuracy, the space of possible entity assignments grows combinatorially. As a consequence, many approaches in this group rely on approximate inference mechanisms. Cucerzan [5] uses high-dimensional vector space representations of candidate entities and attempts to iteratively choose candidates that optimize the mutual proximity to existing candidates. Kulkarni et al. [19] exploit topical information about candidate entities and try to harmonize these topics across all assigned entities. Ratinov et al. [27] prune the list of entity mentions using support vector machines trained on a range of similarity and term overlap features between entity representations. Ferragina and Scaiella [10] focus on short documents such as tweets or search engine snippets. Based on evidence across all mentions, the authors employ a voting scheme for entity disambiguation. Cheng et al. [4] and Singh et al. [31] describe models for jointly capturing the interdependence between the tasks of entity tagging, relation extraction and co-reference resolution. Similarly, Durrett and Klein [8] describe a graphical model for collectively addressing the tasks of named entity recognition, entity disambiguation and co-reference resolution.

Graph-based models establish relationships between candidate entities and mentions using structural models. For inference, various approaches are employed, ranging from densest graph estimation algorithms (Hoffart et al. [15]) to graph traversal methods such as random graph walks (Guo and Barbosa [11], Han et al. [13]). In a similar fashion, these techniques can be combined to enhance the quality of both entity linking and word sense disambiguation in a synergistic solution (Moro et al. [23]).

The above approaches are limited because they assume a single topic per document. Naturally, *topic modelling* can be used for entity disambiguation by attempting to harmonize the individual distribution of latent topics across candidate entities. Houlsby and Ciaramita [16] and Pilz and Paaß [26] rely on Latent Dirichlet Allocation (LDA) and compare the resulting topic distribution of the input document to the topic distributions of the disambiguated entities' Wikipedia pages. Han and Sun [12] propose a joint model of mention context compatibility and topic coherence, allowing them to simultaneously draw from both local (terms, mentions) as well as global (topic distributions) information. Kataria et al. [18] use a semi-supervised hierarchical LDA model based on a wide range of features extracted from Wikipedia pages and topic hierarchies.

In contrast to previous work on this problem, our method exploits co-occurrence statistics in a fully probabilistic manner using a graph-based model that addresses collective entity disambiguation. It combines a clean and light-weight

---

[3]For example, using a named-entity recognition system. However, note that our approach is not restricted to named entities, but targets any Wikipedia entity.

probabilistic model with an elegant, real-time inference algorithm. An advantage over increasingly popular deep learning architectures for entity linking (e.g. Sun et al. [34], He et al. [14]) is the speed of our training procedure that relies on count statistics from data and that learns only very few parameters. State-of-art accuracy is achieved without the need for special-purpose computational heuristics.

3. PROBABILISTIC MODEL

In this section, we formally define the *entity linking* task that we address in this work and describe our modeling approach in detail.

3.1 Problem Definition and Formulation

Let $\mathcal{E}$ be a knowledge base (KB) of entities, $\mathcal{V}$ a finite dictionary of phrases or names and $\mathcal{C}$ a context representation. Formally, we seek a mapping $F : (\mathcal{V}, \mathcal{C})^n \to \mathcal{E}^n$, that takes as input a sequence of linkable *mentions* $\mathbf{m} = (m_1, \ldots, m_n)$ along with their contexts $\mathbf{c} = (c_1, \ldots, c_n)$ and produces a joint entity assignment $\mathbf{e} = (e_1, \ldots, e_n)$. Here n refers to the number of linkable spans in a document. Our problem is also known as *entity disambiguation* or *link generation* in the literature. [4]

We can construct such a mapping F in a probabilistic approach, by learning a conditional probability model $p(\mathbf{e}|\mathbf{m}, \mathbf{c})$ from data and then employing (approximate) probabilistic inference in order to find the maximum a posteriori (MAP) assignment, hence:

$$F(\mathbf{m}, \mathbf{c}) := \arg\max_{\mathbf{e} \in \mathcal{E}^n} p(\mathbf{e}|\mathbf{m}, \mathbf{c}). \quad (1)$$

In the sequel, we describe how to estimate such a model from a corpus of entity-linked documents. Finally, we show in Section 4 how to apply belief propagation (max-product) for approximate inference in this model.

3.2 Maximum Entropy Models

Assume a corpus of entity-linked documents is available. Specifically, we used the set of Wikipedia pages together with their respective Wiki hyperlinks. These hyperlinks are considered ground truth annotations, the mention being the linked span of text and the truth entity being the Wikipedia page it refers to. One can extract two kinds of basic statistics from such a corpus: First, counts of how often each entity was referred to by a specific name. Second, pairwise co-occurrence counts for entities in documents. Our fundamental conjecture is that most of the relevant information needed for entity disambiguation is contained in these counts, that they are *sufficient statistics*. We thus request that our probability model reproduces these counts in expectation. As this alone typically yields an ill-defined problem, we follow the *maximum entropy principle* of Jaynes [17]: Among the feasible set of distributions we favor the one with maximal entropy.

Formally, let $\mathcal{D}$ be an entity-linked document collection. Ignoring mention contexts for now, we extract for each document $d \in \mathcal{D}$ a sequence of mentions $\mathbf{m}^{(d)}$ and their corresponding target entities $\mathbf{e}^{(d)}$, both of length $n^{(d)}$. Assuming exchangeability of random variables within these se-

quences, we reduce each $(\mathbf{e}, \mathbf{m})$ to statistics (or *features*) about mention-entity and entity-entity co-occurrence as follows:

$$\phi_{e,m}(\mathbf{e}, \mathbf{m}) := \sum_{i=1}^{n} \mathbb{1}[e_i = e] \cdot \mathbb{1}[m_i = m], \ \forall (e, m) \in \mathcal{E} \times \mathcal{V} \quad (2)$$

$$\psi_{\{e,e'\}}(\mathbf{e}) := \sum_{i<j} \mathbb{1}[\{e_i, e_j\} = \{e, e'\}], \ \forall e, e' \in \mathcal{E}, \quad (3)$$

where $\mathbb{1}[\cdot]$ is the indicator function. Note that we use the subscript notation $\{e, e'\}$ for ψ to take into account the symmetry in e, e' as well the fact that one may have $e = e'$.

The document collection provides us with empirical estimates for the expectation of these statistics under an i.i.d. sampling model for documents, namely the averages

$$\phi_{e,m}(\mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \phi_{e,m}(\mathbf{e}^{(d)}, \mathbf{m}^{(d)}), \quad (4)$$

$$\psi_{\{e,e'\}}(\mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \psi_{\{e,e'\}}(\mathbf{e}^{(d)}). \quad (5)$$

Note that in entity disambiguation, the mention sequence $\mathbf{m}$ is always considered given, while we seek to predict the corresponding entity sequence $\mathbf{e}$. It is thus not necessary to try to model the joint distribution $p(\mathbf{e}, \mathbf{m})$, but sufficient to construct a conditional model $p(\mathbf{e}|\mathbf{m})$. Following Berger et al. [1] this can be accomplished by taking the empirical distribution $p(\mathbf{m}|\mathcal{D})$ of mention sequences and combining it with a conditional model via $p(\mathbf{e}, \mathbf{m}) = p(\mathbf{e}|\mathbf{m}) \cdot p(\mathbf{m}|\mathcal{D})$. We then require that:

$$\mathbb{E}_p[\phi_{e,m}] = \phi_{e,m}(\mathcal{D}) \quad \text{and} \quad \mathbb{E}_p[\psi_{\{e,e'\}}] = \psi_{\{e,e'\}}(\mathcal{D}), \quad (6)$$

which yields $|\mathcal{E}| \cdot |\mathcal{V}| + \binom{|\mathcal{E}|}{2} + |\mathcal{E}|$ moment constraints on $p(\mathbf{e}|\mathbf{m})$.

The maximum entropy distributions, fulfilling constraints as stated in Eq. (6) form a conditional exponential family for which $\phi(\cdot, \mathbf{m})$ and $\psi(\cdot, \cdot)$ are sufficient statistics. We thus know that there are canonical parameters $\rho_{e,m}$ and $\lambda_{\{e,e'\}}$ (formally corresponding to Lagrange multipliers) such that the maximum entropy distribution can be written as

$$p(\mathbf{e}|\mathbf{m}; \rho, \lambda) = \frac{1}{Z(\mathbf{m})} \exp\left[\langle \rho, \phi(\mathbf{e}, \mathbf{m})\rangle + \langle \lambda, \psi(\mathbf{e})\rangle\right] \quad (7)$$

where $Z(\mathbf{m})$ is the partition function

$$Z(\mathbf{m}) := \sum_{\mathbf{e} \in \mathcal{E}^n} \exp\left[\langle \rho, \phi(\mathbf{e}, \mathbf{m})\rangle + \langle \lambda, \psi(\mathbf{e})\rangle\right]. \quad (8)$$

Here we interpret (e, m) and $\{e, e'\}$ as multi-indices and suggestively define the shorthands

$$\langle \rho, \phi \rangle := \sum_{e,m} \rho_{e,m} \phi_{e,m}, \ \ \langle \lambda, \psi \rangle := \sum_{\{e,e'\}} \lambda_{\{e,e'\}} \psi_{\{e,e'\}}. \quad (9)$$

Note that we can switch between the statistics view and the raw data view by observing that

$$\langle \rho, \phi(\mathbf{e}, \mathbf{m})\rangle = \sum_{i=1}^{n} \rho_{e_i, m_i}, \quad \langle \lambda, \psi(\mathbf{e})\rangle = \sum_{i<j} \lambda_{\{e_i, e_j\}}. \quad (10)$$

While the maximum entropy principle applied to our fundamental conjecture restricts the form of our model to a finite-dimensional exponential family, we need to investigate ways of finding the optimal or – as we will see – an approximately optimal distribution in this family. To that extent, we first re-interpret the obtained model as a factor graph model.
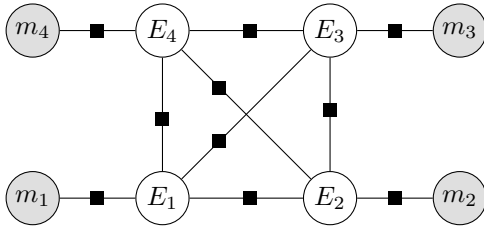
---

[4]Note that we do not address the issues of *mention detection* or *nil identification* in this work. Rather, our input is a document along with a fixed set of linkable mentions corresponding to existing KB entities.

Figure 2: Proposed factor graph for a document with four mentions. Each mention node m_i is paired with its corresponding entity node E_i, while all entity nodes are connected through entity-entity pair factors.

3.3 Markov Network and Factor Graph

Complementary to the maximum entropy estimation perspective, we want to present a view on our model in terms of probabilistic graphical models and factor graphs. Inspecting Eq. (7) and interpreting ϕ and ψ as potential functions, we can recover a Markov network that makes conditional independence assumptions of the following type: an entity link e_i and a mention m_j with $i \neq j$ are independent, given m_i and $\mathbf{e}_{-i}$, where $\mathbf{e}_{-i}$ denotes the set of entity variables in the document excluding e_i. This means that a mention m_j only influences a variable e_i through the intermediate variable e_j. However, the functional form in Eq. (7) goes beyond these conditional independences in that it limits the order of interaction among the variables. A variable e_i interacts with neighbors in its Markov blanket through pairwise potentials. In terms of a *factor graph* decomposition, $p(\mathbf{e}|\mathbf{m})$ decomposes into functions of two arguments only, modeling pairwise interactions between entities on one hand, and between entities and their corresponding mentions on the other hand.

We emphasize the factor model view by rewriting (7) as

$$p(\mathbf{e}|\mathbf{m}; \rho, \lambda) \propto \prod_i \exp\left[\rho_{e_i, m_i}\right] \cdot \prod_{i<j} \exp\left[\lambda_{\{e_i, e_j\}}\right] \quad (11)$$

where we think of ρ and λ as functions

$$\rho : \mathcal{E} \times \mathcal{V} \to \mathbb{R}, \quad (e, m) \mapsto \rho_{e,m}$$

$$\lambda : \mathcal{E} \cup \mathcal{E}^2 \to \mathbb{R}, \quad \{e, e'\} \mapsto \lambda_{\{e,e'\}}$$

An example of a factor graph ($n = 4$) is shown in Figure 2. We will investigate in the sequel how the factor graph structure can be further exploited.

3.4 (Pseudo–)Likelihood Maximization

While the maximum entropy approach directly motivates the exponential form of Eq. (7) and is amenable to a plausible factor graph interpretation, it does not by itself suggest an efficient parameter fitting algorithm. As is known by convex duality, the optimal parameters can be obtained by maximizing the conditional likelihood of the model under the data,

$$\mathcal{L}(\rho, \lambda; \mathcal{D}) = \sum_d \log\ p(\mathbf{e}^{(d)}|\mathbf{m}^{(d)}; \rho, \lambda) \quad (12)$$

However, specialized algorithms for maximum entropy estimation such as generalized iterative scaling [6] are known to be slow, whereas gradient-based methods require the com-

putation of gradients of $\mathcal{L}$, which involves evaluating expectations with regard to the model, since

$$\nabla_\rho \log Z(\mathbf{m}) = \mathbb{E}_p \phi(\mathbf{e}, \mathbf{m}), \quad \nabla_\lambda \log Z(\mathbf{m}) = \mathbb{E}_p \psi(\mathbf{e}). \quad (13)$$

The exact inference problem of computing these model expectations, however, is not generally tractable due to the pairwise couplings through the ψ-statistics.

As an alternative to maximizing the likelihood in Eq. (12), we have investigated an approximation known as, pseudo-likelihood maximization [35, 38]. Its main benefits are low computational complexity, simplicity and practical success. Switching to the Markov network view, the pseudo-likelihood estimator predicts each variable conditioned on the value of all variables in its Markov blanket. The latter consists of the minimal set of variables that renders a variable conditionally independent of everything else. In our case the Markov blanket consists of all variables that share a factor with a given variable. Consequently, the Markov blanket of e_i is $\mathcal{N}(e_i) := (m_i, \mathbf{e}_{-i})$. The posterior is then approximated in the pseudo-likelihood approach as:

$$\tilde{p}(\mathbf{e}|\mathbf{m}; \rho, \lambda) := \prod_{i=1}^{n} p(e_i|\mathcal{N}(e_i); \rho, \lambda), \quad (14)$$

which results in the tractable log-likelihood function

$$\tilde{\mathcal{L}}(\rho, \lambda; \mathcal{D}) := \sum_{d \in \mathcal{D}} \sum_{i=1}^{n^{(d)}} \log p(e_i^{(d)}|\mathcal{N}(e_i^{(d)}); \rho, \lambda). \quad (15)$$

Introducing additional L_2-norm penalties $\gamma(\|\lambda\|_2^2 + \|\rho\|_2^2)$ to further regularize $\tilde{\mathcal{L}}$, we have utilized parallel stochastic gradient descent (SGD) [28] with sparse updates to learn parameters ρ, λ. From a practical perspective, we only keep for each token span m parameters $\rho_{e,m}$ for the most frequently observed entities e. Moreover, we only use $\lambda_{\{e,e'\}}$ for entity pairs (e, e') that co-occurred together a sufficient number of times in the collection $\mathcal{D}$.[5] As we will discuss in more detail in Section 5, our experimental findings suggest this brute-force learning approach to be somewhat ineffective, which has motivated us to develop simpler, yet more effective plug-in estimators as described below.

3.5 Bethe Approximation

The major computational difficulty with our model lies in the pairwise couplings between entities and the fact that these couplings are dense: The Markov dependency graph between different entity links in a document is always a complete graph. Let us consider what would happen, if the dependency structure were loop-free, i.e., it would form a tree. Then we could rewrite the prior probability in terms of marginal distributions in the so-called *Bethe form*. Encoding the tree structure in a symmetric relation $\mathcal{T}$, we would get

$$p(\mathbf{e}) = \frac{\prod_{\{i,j\} \in \mathcal{T}} p(e_i, e_j)}{\prod_{i=1}^{n} p(e_i)^{d_i - 1}}, \quad d_i := |\{j : \{i,j\} \in \mathcal{T}\}|. \quad (16)$$

The Bethe approximation [39] pursues the idea of using the above representation as an unnormalized approximation for

---

[5]For the Wikipedia collection, even after these pruning steps, we ended up with more than 50 million parameters in total.

$p(\mathbf{e})$, even when the Markov network has cycles. How does this relate to the exponential form in Eq. (7)? By simple pattern matching, we see that if we choose

$$\lambda_{\{e,e'\}} = \log\left(\frac{p(e,e')}{p(e)\,p(e')}\right), \quad \forall e,e' \in \mathcal{E} \qquad (17)$$

we can apply Eq. (16) to get an approximate distribution

$$
\bar{p}(\mathbf{e}) \propto \frac{\prod_{i<j} p(e_i, e_j)}{\prod_{i=1}^{n} p(e_i)^{n-2}} = \prod_{i=1}^{n} p(e_i) \prod_{i<j} \frac{p(e_i, e_j)}{p(e_i)\,p(e_j)}
$$
$$
= \exp\left[\sum_i \log p(e_i) + \sum_{i<j} \lambda_{\{e_i, e_j\}}\right], \qquad (18)
$$

where we see the same exponential form in λ appearing as in Eq. (10). We complete this argument by observing that with

$$\rho_{e,m} = \log p(e) + \log p(m|e) \qquad (19)$$

we obtain a representation of a joint distribution that exactly matches the form in Eq. (7).

What have we gained so far? We started from the desire of constructing a model that would agree with the observed data on the co-occurrence probabilities of token spans and their linked entities as well as on the co-link probability of entity pairs within a document. This has led to the conditional exponential family in Eq. (7). We have then proposed pseudo-likelihood maximization as a way to arrive at a tractable learning algorithm to try to fit the massive amount of parameters ρ and λ. Alternatively, we have now seen that a Bethe approximation of the joint prior $p(\mathbf{e})$ yields a conditional distribution $p(\mathbf{e}|\mathbf{m})$ that (i) is a member of the same exponential family, (ii) has explicit formulas for how to choose the parameters from pairwise marginals, and (iii) would be exact in the case of a dependency tree. We claim that the benefits of computational simplicity together with the correctness guarantee for non-dense dependency networks outweighs the approximation loss, relative to the model with the best generalization performance within the conditional exponential family. In order to close the suboptimality gap further, we suggest some important refinements below.

3.6 Parameter Calibration

With the previous suggestion, one issue comes into play: The total contribution coming from the pairwise interactions between entities will scale with $\binom{n}{2}$, while the entity–mention compatibility contributions will scale with n, the total number of mentions. This is a direct observation of the number of terms contributing to the sums in (10). However, for practical reasons, it is somewhat implausible that, as n grows, the prior $p(\mathbf{e})$ should dominate and the contribution of the likelihood term should vanish. The model is not well-calibrated with regard to n.

We propose to correct for this effect by adding a normalization factor to the λ-parameters by replacing (17) with:

$$\lambda_{e,e'}^{n} = \frac{2}{n-1} \log\left(\frac{p(e,e')}{p(e)\cdot p(e')}\right), \quad \forall e,e' \in \mathcal{E} \qquad (20)$$

where now these parameters scale inversely with n, the number of entity links in a document, making the corresponding

sum in Eq. (7) scale with n. With this simple change, a substantial accuracy improvement was observed empirically, the details of which are reported in our experiments.

The re-calibration in Eq. (20) can also be justified by the following combinatorial argument: For a given set $\mathbf{Y}$ of random variables, define an $\mathbf{Y}$-cycle as a graph containing as nodes all variables in $\mathbf{Y}$, each with degree exactly 2, connected in a single cycle. Let Ξ be the set enumerating all possible $\mathbf{Y}$-cycles. Then, $|\Xi| = (n-1)!$, where n is the size of $\mathbf{Y}$.

In our case, if the entity variables $\mathbf{e}$ per document would have formed a cycle of length n instead of a complete subgraph, the Bethe approximation would have been written as:

$$\bar{p}_\pi(\mathbf{e}) \propto \frac{\prod_{(i,j)\in E(\pi)} p(e_i, e_j)}{\prod_i p(e_i)}, \quad \forall \pi \in \Xi \qquad (21)$$

where $E(\pi)$ is the set of edges of the $\mathbf{e}$-cycle π. However, as we do not desire to further constrain our graph with additional independence assumptions, we propose to approximate the joint prior $p(\mathbf{e})$ by the average of the Bethe approximation of all possible π, that is

$$\log \bar{p}(\mathbf{e}) \approx \frac{1}{|\Xi|} \sum_{\pi \in \Xi} \log \bar{p}_\pi(\mathbf{e}). \qquad (22)$$

Since each pair (e_i, e_j) would appear in exactly $2(n-2)!$ $\mathbf{e}$-cycles, one can derive the final approximation:

$$\bar{p}(\mathbf{e}) \approx \frac{\prod_{i<j} p(e_i, e_j)^{\frac{2}{n-1}}}{\prod_i p(e_i)}. \qquad (23)$$

Distributing marginal probabilities over the parameters starting from Eq. (23) and applying a similar argument as in Eq. (18) results in the assignment given by Eq. (20). While the above line of argument is not a strict mathematical derivation, we believe this to shed further light on the empirically observed effectiveness of the parameter re-scaling.

3.7 Integrating Context

The model that we have discussed so far does not consider the local context of a mention. This is a powerful source of information that a competitive entity linking system should utilize. For example, words like "computer", "company" or "device" are more likely to appear near references of the entity Apple_Inc. than of the entity Apple_fruit. We demonstrate in this section how this integration can be easily done in a principled way on top of the current probabilistic model. This showcases the extensibility of our approach. Enhancing our model with additional knowledge such as entity categories or word co-reference can also be done in a rigorous way, so we hope that this provides a template for future extensions.

As stated in Section 3.1, for each mention m_i in a document, we maintain a context representation c_i consisting of the bag of words surrounding the mention within a window of length K^6. Hence, c_i can be viewed as an additional random variable with an observed outcome. At this stage, we make additional reasonable independence assumptions that increase tractability of our model. First, we assume

---

[6]Throughout our experiments, we used a context window of size $K = 100$, intuitively chosen and without extensive validation.

that, knowing the identity of the linked entity e_i, the mention token span m_i is just the surface form of the entity, so it brings no additional information for the generative process describing the surrounding context c_i. Formally, this means that m_i and c_i are conditionally independent given e_i. Consequently, we obtain a factorial expression for the joint model

$$p(\mathbf{e}, \mathbf{m}, \mathbf{c}) = p(\mathbf{e})p(\mathbf{m}, \mathbf{c}|\mathbf{e}) = p(\mathbf{e}) \prod_{i=1}^{n} p(m_i|e_i)p(c_i|e_i) \quad (24)$$

This is a simple extension of the previous factor graph that includes context variables. Second, we assume conditional independence of the words in c_i given an entity e_i which let us factorize the context probabilities as

$$p(c_i|e_i) = \prod_{w_j \in c_i} p(w_j|e_i). \quad (25)$$

Note that this assumption is commonly made in models using bag-of-word representations or naïve Bayes classifiers.

While this completes the argument from a joint model point of view, we need to consider one more aspect for the conditional distribution $p(\mathbf{e}|\mathbf{m}, \mathbf{c})$ that we are interested in. If we cannot afford (computationally as well as with regard to training data size) a full-blown discriminative learning approach, then how do we balance the relative influence of the context c_i and the mention token span m_i on e_i? For instance, the effect of c_i will depend on the chosen window size K, which is not realistic.

To address this issue, we resort to a hybrid approach, where, in the spirit of the Bethe approximation, we continue to express our model in terms of simple marginal distributions that can be easily estimated independently from data, yet that allow for a small number of parameters (in our case "small" equals 2) to be chosen to optimize the conditional log-likelihood $p(\mathbf{e}|\mathbf{m}, \mathbf{c})$. We thus introduce weights ζ and τ that control the importance of the context factors and, respectively, of the entity-entity interaction factors. Putting equations (19), (20), (24) and (25) together, we arrive at the final model that will be subsequently referred to as the *PBoH* model (**P**robabilistic **B**ag **o**f **H**yperlinks):

$$\log p(\mathbf{e}|\mathbf{m}, \mathbf{c}) = \sum_{i=1}^{n} \left(\log p(e_i|m_i) + \zeta \sum_{w_j \in c_i} \log p(w_j|e_i) \right)$$
$$+ \frac{2\tau}{n-1} \sum_{i<j} \log \left(\frac{p(e_i, e_j)}{p(e_i)\, p(e_j)} \right) + \text{const}. \quad (26)$$

Here we used the identity $p(m|e)p(e) = p(e|m)p(m)$ and absorbed all $\log p(m)$ terms in the constant. We use grid-search on a validation set for the remaining problem of optimizing over the parameters ζ, τ. Details are provided in section 5.

3.8 Smoothing Empirical Probabilities

In order to estimate the probabilities involved in Eq. (26), we rely on an entity annotated corpus of text documents, e.g., Wikipedia Web pages together with their hyperlinks which we view as ground truth annotations. From this corpus, we derive empirical probabilities for a name-to-entity dictionary $\hat{p}(m|e)$ based on counting how many times an en-

tity appeared referenced by a given name[7]. We also compute the pairwise probabilities $\hat{p}(e, e')$ obtained by counting the pairwise co-occurrence of entities e and e' within the same document. Similarly, we obtained empirical values for the marginals $\hat{p}(e) = \sum_{e'} \hat{p}(e, e')$ and for the context word-entity statistics $\hat{p}(w|e)$.

In the absence of huge amounts of data, estimating such probabilities from counts is subject to sparsity. For instance, in our statistics, there are 8 times more distinct pairs of entities that co-occur in at most 3 Wikipedia documents compared to the total number of distinct pairs of entities that appear together in at least 4 documents. Thus, it is expected that the heavy tail of infrequent pairs of entities will have a strong impact on the accuracy of our system.

Traditionally, various smoothing techniques are employed to address sparsity issues arising commonly in areas such as natural language processing. Out of the wealth of methods, we decided to use the absolute discounting smoothing technique [40] that involves interpolation of higher and lower order (backoff) models. In our case, whenever insufficient data is available for a pair of entities (e, e'), we assume the two entities are drawn from independent distributions. Thus, if we denote by $N(e, e')$ the total number of corpus documents that link both e and e', and by N_{ep} the total number of pairs of entities referenced in each document, then the final formula for the smoothed entity pairwise probabilities is:

$$\widetilde{p}(e, e') = \frac{\max(N(e, e') - \delta, 0)}{N_{ep}} + (1 - \mu_e)\hat{p}(e)\hat{p}(e') \quad (27)$$

where $\delta \in [0, 1]$ is a fixed discount and μ_e is a constant that assures that $\sum_e \sum_{e'} \widetilde{p}(e, e') = 1$. δ was set by performing a coarse grid search on a validation set. The best δ value was found to be 0.5.

The word-entity empirical probabilities $\hat{p}(w|e)$ were computed based on the Wikipedia corpus by counting the frequency with which word w appears in the context windows of size K around the hyperlinks pointing to e. In order to avoid memory explosion, we only considered the entity-words pairs for which these counts are at least 3. These empirical estimates are also sparse, so we used absolute discounting smoothing for their correction by backing off to the unbiased estimates $\hat{p}(w)$. The latter can be much more accurately estimated from any text corpus. Finally, we obtain:

$$\widetilde{p}(w|e) = \frac{\max(N(w, e) - \xi, 0)}{N_{wp}} + (1 - \mu_w)\hat{p}(w). \quad (28)$$

Again $\xi \in [0, 1]$ was optimized by grid search to be 0.5.

4. INFERENCE

After introducing our model and showing how to train it in the previous section, we now explain the inference process used for prediction.

4.1 Candidate Selection

At test time, for each mention to be disambiguated, we first select a set of potential candidates by considering the top R ranked entities based on the local mention-entity probability dictionary $\hat{p}(e|m)$. We found $R = 64$ to be a good

---

[7]In our implementation we summed the mention-entity counts from Wikipedia hyperlinks with the Crosswikis counts [32]

compromise between efficiency and accuracy loss. Second, we want to keep the average number of candidates per mention as small as possible in order to reduce the running time which is quadratic in this number (see the next section for details). Consequently, we further limit the number of candidates per mention by keeping only the top 10 entity candidates re-ranked by the local mention-context-entity compatibility defined as

$$\log p(e_i | m_i, c_i) = \log p(e_i | m_i) + \zeta \sum_{w_j \in c_i} \log p(w_j | e_i) + \text{const} . \tag{29}$$

These pruning heuristics result in a significantly improved running time at an insignificant accuracy loss.

If the given mention is not found in our map $\hat{p}(e|m)$, we try to replace it by the closest name in this dictionary. Such a name is picked only if the Jaccard distance between the set of letter trigrams of these two strings is smaller than a threshold that we empirically picked as 0.5. Otherwise, the mention is not linked at all.

4.2 Belief Propagation

Collectively disambiguating all mentions in a text involves iterating through an exponential number of possible entity resolutions. Exact inference in general graphical models is NP-hard, therefore approximations are employed. We propose solving the inference problem through the *loopy belief propagation* (LBP) [24] technique, using the max-product algorithm that approximates the MAP solution in a run-time polynomial in n, the number of input mentions. For the sake of brevity, we only present the algorithm for the maximum entropy model described by Eq. (7); A similar approach was used for the enhanced PBoH model given by Eq. (26).

Our proposed graphical model is a fully connected graph where each node corresponds to an entity random variable. Unary potentials $\exp(\rho_{m,e})$ model the entity-mention compatibility, while pairwise potentials $\exp(\lambda_{\{e,e'\}})$ express entity-entity correlations. For the posterior in Eq. (7), one can derive the update equation of the logarithmic message that is sent in round $t + 1$ from entity random variable E_i to the outcome e_j of the entity random variable E_j :

$$m_{E_i \to E_j}^{t+1}(e_j) = \tag{30}$$

$$\max_{e_i} \left(\rho_{e_i, m_i} + \lambda_{\{e_i, e_j\}} + \sum_{1 \leq k \leq n; k \neq j} m_{E_k \to E_i}^{t}(e_i) \right)$$

Note that, for simplicity, we skip the factor graph framework and send messages directly between each pair of entity variables. This is equivalent to the original BP framework.

We chose to update messages synchronously: in each round t, each two entity nodes E_i and E_j exchange messages. This is done until convergence or until an allowed maximum number of iterations (15 in our experiments) is reached. The convergence criterion is:

$$\max_{1 \leq i,j \leq n; e_j \in \mathcal{E}} | m_{E_i \to E_j}^{t+1}(e_j) - m_{E_i \to E_j}^{t}(e_j) | \leq \epsilon \tag{31}$$

where $\epsilon = 10^{-5}$. This setting was sufficient in most of the cases to reach convergence.

Dataset	# non-NIL mentions	# documents
AIDA test A	4791	216
AIDA test B	4485	231
MSNBC	656	20
AQUAINT	727	50
ACE04	257	35

Table 1: Statistics on some of the used datasets

In the end, the final entity assignment is determined by:

$$e_i^* = \arg \max_{e_i} \left(\rho_{e_i, m_i} + \sum_{1 \leq k \leq n} m_{E_k \to E_i}^{t}(e_i) \right) \tag{32}$$

The complexity of the belief propagation algorithm is, in our case, $O(n^2 \cdot r^2)$, with n being the number of mentions in a document and r being the average number of candidate entities per mention (10 in our case). More details regarding the run-time and convergence of the loopy BP algorithm can be found in Section 5.

5. EXPERIMENTS

We now present the experimental evaluation of our method. We first uncover some practical details of our approach. Further, we show an empirical comparison between PBoH and well known or recent competitive entity disambiguation systems. We use the Gerbil testing platform [37] version 1.1.4 with the D2KB setting in which a document together with a fixed set of mentions to be annotated are given as input. We run additional experiments that allow us to compare against more recent approaches, such as [16] and [11].

Note that in all the experiments we assume that we have access to a set of linkable token spans for each document. In practice this set is obtained by first applying a mention detection approach which is not part of our method. Our main goal is then to annotate each token span with a Wikipedia entity[8].

Evaluation metrics. We quantify the quality of an entity linking system by measuring common metrics such as **precision**, **recall** and **F₁ scores**.

Let M^* be the ground truth entity annotations associated with a given set of mentions X. Note that in all the results reported, mentions that contain NIL or empty ground truth entities are discarded before the evaluation; this decision is taken as well in Gerbil version 1.1.4. Let M be the output annotations of an entity disambiguation system on the same input. Then, our quality metrics are computed as follows:

- Precision: $P = \frac{|M \cap M^*|}{|M|}$

- Recall: $R = \frac{|M \cap M^*|}{|M^*|}$

- F_1 score: $F_1 = \frac{2 \cdot P \cdot R}{P + R}$

We mostly report results in terms of F_1 scores, namely macro-averaged **F1@MA** (aggregated across documents), and micro-averaged **F1@MI** (aggregated across mentions). For a fair comparison with Houlsby and Ciaramita [16], we

---

[8]In PBoH, we refrain from annotating mentions for which no candidate entity is found according to the procedure described in Section 4.1.

F1@MI / F1@MA	ACE2004	AIDA/CoNLL-Complete	AIDA/CoNLL-Test A	AIDA/CoNLL-Test B	AIDA/CoNLL-Training	AQUAINT	DBpediaSpotlight	IITB	KORE50	Microposts2014-Test	Microposts2014-Train	MSNBC	N3-Reuters-128	N3-RSS-500
AGDISTIS	65.83	60.27	59.06	58.32	61.05	60.10	36.61	41.23	34.16	42.43	50.39	75.42	67.95	59.88
	77.63	56.97	53.36	58.03	57.53	58.62	33.25	43.38	30.20	61.08	62.87	73.82	75.52	70.80
Babelfy	63.20	78.00	75.77	80.36	78.01	72.27	51.05	57.13	73.12	47.20	50.60	78.17	58.61	69.17
	76.71	73.81	71.26	74.52	74.22	73.23	51.97	55.36	69.77	62.11	61.02	75.73	59.87	76.00
DBpedia Spotlight	70.38	58.84	54.90	57.69	60.04	74.03	69.27	65.44	37.59	56.43	56.26	69.27	56.44	57.63
	80.02	60.59	54.11	61.34	62.23	73.13	67.23	62.81	32.90	71.63	67.99	69.82	58.77	65.03
Dexter	18.72	48.46	45.44	48.59	49.25	38.28	26.70	28.53	17.20	31.27	35.21	36.86	32.74	31.11
	16.97	45.29	42.17	46.20	45.85	38.15	22.75	28.48	12.54	44.02	42.07	39.42	31.85	33.55
Entityclassifier.eu	12.74	46.6	44.13	44.02	47.83	21.67	22.59	18.46	27.97	29.12	32.69	41.24	28.4	21.77
	12.3	42.86	42.36	41.31	43.36	19.59	18.0	19.54	25.2	39.53	38.41	40.3	24.84	22.2
Kea	80.08	73.39	70.9	72.64	74.22	81.84	73.63	72.03	57.95	63.4	64.67	85.49	63.2	69.29
	87.57	73.26	67.91	73.31	74.47	81.27	76.60	70.52	53.17	76.54	74.32	87.4	64.45	75.93
NERD-ML	54.89	54.62	52.85	52.59	55.55	49.68	46.8	51.08	29.96	38.65	39.83	64.03	54.96	61.22
	72.22	52.35	49.6	51.34	53.23	46.06	45.59	49.91	24.75	57.91	53.74	67.28	62.9	67.3
TagMe 2	81.93	72.07	69.07	70.62	73.2	76.27	63.31	57.23	57.34	56.81	59.14	75.96	59.32	78.05
	89.09	71.19	66.5	70.38	72.45	75.12	65.1	55.8	54.67	71.66	70.45	77.05	67.55	83.2
WAT	80.0	83.82	81.82	84.34	84.21	76.82	65.18	61.14	58.99	59.56	61.96	77.72	64.38	68.21
	86.49	83.59	80.25	84.12	84.22	77.64	68.24	59.36	53.13	73.89	72.65	79.08	65.81	76.0
Wikipedia Miner	77.14	64.72	61.65	60.71	66.48	75.96	62.57	58.59	41.63	54.88	55.93	64.25	60.05	64.54
	86.36	66.17	61.67	63.19	67.93	74.63	61.43	56.98	35.0	69.29	67.0	64.68	66.51	72.23
PBoH	87.19	86.72	86.63	87.39	86.59	86.64	79.48	62.47	61.70	74.19	73.08	89.54	76.54	71.24
	90.40	86.85	85.48	86.32	87.30	86.14	80.13	61.04	55.83	84.48	81.25	89.62	83.31	78.33

Table 2: **Micro and macro F1 scores reported by Gerbil for 14 datasets and 11 entity linking systems including PBoH. For each dataset and each metric, we highlight in red the best system and in blue the second best system.**

Systems	Datasets			
	AIDA test A		AIDA test B	
	R@MI	R@MA	R@MI	R@MA
LocalMention	69.73	69.30	67.98	72.75
TagMe reimpl.	76.89	74.57	78.64	78.21
AIDA	79.29	77.00	82.54	81.66
S & Y	-	84.22	-	-
Houlsby et al.	79.65	76.61	84.89	83.51
PBoH	85.70	85.26	87.61	86.44

Table 3: AIDA test-a and AIDA test-b datasets results.

also report micro-recall **R@MI** and macro-recall **R@MA** on the AIDA datasets.

Note that, in our case, the precision and recall are not necessarily identical since a method may not consider annotating certain mentions [8].

Pseudo-likelihood training. We briefly mention some of the practical issues that we encounter with the likelihood maximization described in Section 3.4. From the practical perspective, for each mention m, we only considered the set of parameters $\rho_{m,e}$ limited to the top 64 candidate entities e per mention, ranked by $\hat{p}(e|m)$. Additionally, we restricted the set $\lambda_{e,e'}$ to entity pairs (e, e') that co-occurred together in at least 7 documents throughout the Wikipedia corpus. In total, a set of 26 millions ρ and 39 millions λ parameters were

learned using the previously described procedure. Note that the universe of all Wikipedia entities is of size ~ 4 million.

For the SGD procedure, we tried different initializations of these parameters, including $\rho_{m,e} = \log\ p(e|m), \lambda_{e,e'} = 0$, as well as the parameters given by Eq. (17). However, in all cases, the accuracy gain on a sample of 1000 Wikipedia test pages was small or negligible compared to the LocalMention baseline (described below). One reason is the inherent sparsity of the data: the parameters associated with the long tail of infrequent entity pairs are updated rarely and expected to be defective at the end of the SGD procedure. However, these scattered pairs are crucial for the effectiveness and coverage of the entity disambiguation system. To overcome this problem, we refined our model as described in Section 3.5 and subsequent sections.

PBoH training details. Wikipedia itself is a valuable resource for entity linking since each internal hyperlink can be considered as the ground truth annotation for the respective anchor text. In our system, the training is solely done on the entire Wikipedia corpus[9]. Hyper-parameters are grid-searched such that the micro F_1 plus macro F_1 scores are maximized over the combined held-out set containing only the *AIDA Test-A* dataset and a Wikipedia validation set consisting of random 1000 pages. As a preprocessing step in our training procedure, we removed all annotations and hyperlinks that point to non-existing, disambiguation or list Wikipedia pages.

---

[9]We used the Wikipedia dump from February 2014

Systems	Datasets					
	new MSNBC		new AQUAINT		new ACE2004	
	F1@MI	F1@MA	F1@MI	F1@MA	F1@MI	F1@MA
LocalMention	73.64	77.71	87.33	86.80	84.75	85.70
Cucerzan	88.34	87.76	78.67	78.22	79.30	78.22
M & W	78.43	80.37	85.13	84.84	81.29	84.25
Han et al.	88.46	87.93	79.46	78.80	73.48	66.80
AIDA	78.81	76.26	56.47	56.46	80.49	84.13
GLOW	75.37	77.33	83.14	82.97	81.91	83.18
RI	90.22	90.87	87.72	87.74	86.60	87.13
REL-RW	91.37	91.73	90.74	90.58	87.68	89.23
PBoH	91.06	91.19	89.27	88.94	88.71	88.46

Table 4: Results on the newer versions of the MSNBC, AQUAINT and ACE04 datasets.

The *PBoH* system used in the experimental comparison is the model given by Eq. (26) for which grid search of the hyper-parameters suggested using $\zeta = 0.075, \tau = 0.5, \delta = 0.5, \xi = 0.5$.

Datasets. We evaluate our approach on 14 well-known public entity linking datasets built from various sources. Statistics of some of them are shown in Table 1, and their descriptions are provided below. For information on the other datasets used only in the Gerbil experiments, refer to [37].

- The *CoNLL-AIDA* dataset is an entity annotated corpus of Reuters news documents introduced by Hoffart et al. [15]. It is much larger than most of the other existing EL datasets, making it an excellent evaluation target. The data is divided in three parts: Train (not used in our current setting for training, but only in the Gerbil evaluation), Test-A (used for validation) and Test-B (used for blind evaluation). Similar to Houlsby and Ciaramita [16] and others, we report results also on the validation set Test-A.

- The *AQUAINT* dataset introduced by Milne and Witten [22] contains documents from a news corpus from the Xinhua News Service, the New York Times and the Associated Press.

- *MSNBC* [5] - a dataset of news documents that includes many mentions which do not easily map to Wikipedia titles because of their rare surface forms or distinctive lexicalization.

- The *ACE04* dataset [27] is a subset of ACE2004 Coreference documents annotated using Amazon Mechanical Turk. Note that the ACE04 dataset contains mentions that are annotated with NIL entities, meaning that no proper Wikipedia entity was found. Following common practice, we removed all the mentions corresponding to these NIL entities prior to our evaluation.

Note that the Gerbil platform uses an old version of the AQUAINT, MSNBC and ACE04 datasets that contain some no-longer existing Wikipedia entities. A new cleaned version of these sets[10] was released by Guo & Barbosa [11]. We report results for the new cleaned datasets in Table 4, while Table 2 contains results for the old versions currently used by Gerbil.

---

[10]http://www.cs.ualberta.ca/~denilson/data/
deos14_ualberta_experiments.tgz

	Datasets				
	AIDA test A	AIDA test B	MSNBC	AQUAINT	ACE04
Avg. num mentions per doc	22.18	19.41	32.8	14.54	7.34
Conv. rate	100%	99.56%	100%	100%	100%
Avg. running time (ms/doc)	445.56	203.66	371.65	40.42	10.88
Avg. num. rounds	2.86	2.83	3.0	2.56	2.25

Table 5: Loopy belief propagation statistics. Average running time, number of rounds and convergence rate of our inference procedure are provided.

Systems. For comparison, we selected a broad range of competitor systems from the vast literature in this field. The Gerbil platform already integrates the methods of Agdistis [36], Babelfy [23], DBpedia Spotlight [20], Dexter [3], Kea [33], Nerd-ML [29], Tagme2 [9], WAT [25], Wikipedia Miner [22] and Illinois Wikifier [27]. We furthermore compare against Cucerzan [5] – the first collective EL system that uses optimization techniques, M& W [22]– a popular machine learning approach, Han et al. [13] – a graph based disambiguation system that uses random walks for joint disambiguation, AIDA [15] – a performant graph based approach, GLOW [27] – a system that uses local and global context to perform joint entity disambiguation, RI [4] – an approach using relational inference for mention disambiguation, and REL-RW [11], a recent system that iteratively solves mentions relying on an online updating random walk model. In addition, on the AIDA datasets we also compare against S& Y [30] – an apparatus for combining the NER and EL tasks, and Houlsby et al. [16] – a topic modelling LDA-based approach for EL.

To empirically assess the accuracy gain introduced by each incremental step of our approach, we ran experiments on several of our method's components, individually: **Local-Mention** – links mentions to entities solely based on the token span statistics, i.e., $e^* = \arg\max_e \hat{p}(e|m)$; **Unnorm** – uses the unnormalized mention-entity model described in Section 3.5; **Rescaled** – relies on the rescaled model presented in Section 3.6; **LocalContext** – disambiguates an entity based on the mention and the local context probability given by Equation (29), i.e., $e^* = \arg\max_e p(e|m,c)$. Note that **Unnorm**, **Rescaled** and **PBoH** use the loopy belief propagation procedure for inference.

935

	Datasets		
	MSNBC	AQUAINT	ACE2004
Avg # mentions per doc	36.95	14.54	8.68
Systems	# entities	# entities	# entities
PBoH	247.19	95.38	66.66
REL-RW	382773.6	242443.1	256235.49

Table 6: Average number of entities that appear in the graph built by PBoH and by REL-RW

5.1 Results

Results of the experiments run on the Gerbil platform are shown in Table 2. Detailed results are also provided[1112]. We obtain the highest performance on 11 datasets and the second highest performance on 2 datasets, showing the effectiveness of our method.

Other results are presented in Table 3 and Table 4. The highest accuracy for the cleaned version of AQUAINT, MSNBC and ACE04 was previously reported by Guo & Barbosa [11], while Houlsby et al. [16] dominate the AIDA datasets. Note that the performance of the baseline systems shown in these two tables is taken from [11] and [16].

All these methods are tested in the setting where a fixed set of mentions is given as input, without requiring the mention detection step.

Discussion. Several observations are worth noting here. First, the simple *LocalMention* component alone outperforms many EL systems. However, our experimental results show that *PBoH* consistently beats *LocalMention* on all the datasets. Second, *PBoH* produces state-of-the-art results on both development (Test-A) and blind evaluation (Test-B) parts of the AIDA dataset. Third, on the AQUAINT, MSNBC and ACE04 datasets, *PBoH* outperforms all but one of the presented EL systems and is competitive with the state-of-art approaches. The method whose performance is closer to ours is REL-RW [11] whose average F1 score is only slightly higher than ours (*+0.6 on average*). However, there are significant advantages of our method that make it easier to use for practitioners. First, our approach is conceptually simpler and only requires sufficient statistics computed from Wikipedia. Second, *PBoH* shows a superior computational complexity manifested in significantly lower run times (Table 5), making it a good fit for large-scale real-time entity linking systems; this is not the case for REL-RW qualified as "time consuming" by its authors. Third, the number of entities in the underlying graph, and thus the required memory, is significantly lower for *PBoH* (see statistics provided in Table 6).

Incremental accuracy gains

To give further insight to our method, Table 7 provides an overview of the contribution brought step by step by each incremental component of the *Full PBoH* system. It can be noted that *PBoH* performs best, outranking all its individual components.

---

[11]The PBoH Gerbil experiment is available at `http://gerbil.aksw.org/gerbil/experiment?id=201510160025`.

[12]The detailed Gerbil results of the baseline systems can be accessed at `http://gerbil.aksw.org/gerbil/experiment?id=201510160026`

	Datasets			
	AIDA test A		AIDA test B	
Systems	**R@MI**	**R@MA**	**R@MI**	**R@MA**
LocalMention	69.73	69.30	67.98	72.75
Unnorm	69.77	69.95	75.87	75.12
Rescaled	75.09	74.25	74.76	78.28
LocalContext	82.50	81.56	85.46	84.08
PBoH	**85.53**	**85.09**	**87.51**	**86.39**

Table 7: Accuracy gains of individual *PBoH* components.

Reproducibility of the experiments

Our experiments are easily reproducible using the details provided in this paper. Our learning procedure is only based on statistics coming from the set of Wikipedia webpages. As a consequence, one can implement a real-time highly accurate entity disambiguation system solely based on the details described in this paper.

Our code is publicly available at : `https://github.com/dalab/pboh-entity-linking`

6. CONCLUSION

In this paper, we described a light-weight graphical model for entity linking via approximate inference. Our method employs simple sufficient statistics that rely on three sources of information: First, a probabilistic name to entity map $\hat{p}(e|m)$ derived from a large corpus of hyperlinks; second, observational data about the pairwise co-occurrence of entities within documents from a Web collection; third, entity - contextual words statistics. Our experiments based on a number of popular entity linking benchmarking collections show improved performance as compared to several well-known or recent systems.

There are several promising directions of future work. Currently, our model considers only pairwise potentials. In the future, it would be interesting to investigate the use of higher-order potentials and submodular optimization in an entity linking pipeline, thus allowing us to capture the interplay between entire groups of entity candidates (e.g., through the use of entity categories). Additionally, we will further enrich our probabilistic model with statistics from new sources of information. We expect some of the performance gains that other papers report from using entity categories or semantic relations to be additive with regard to our system's current accuracy.

7. REFERENCES

[1] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[2] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, pages 9–16, 2006.

[3] D. Ceccarelli, C. Lucchese, S. Orlando, R. Perego, and S. Trani. Dexter: an open source framework for entity linking. In *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*, pages 17–20. ACM, 2013.

[4] X. Cheng and D. Roth. Relational inference for wikification. *Urbana*, 51:61801, 2013.

[5] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, volume 7, pages 708–716, 2007.

[6] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The annals of mathematical statistics*, pages 1470–1480, 1972.

[7] M. Dredze, P. McNamee, D. Rao, A. Gerber, and T. Finin. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 277–285. Association for Computational Linguistics, 2010.

[8] G. Durrett and D. Klein. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490, 2014.

[9] P. Ferragina and U. Scaiella. Fast and accurate annotation of short texts with wikipedia pages. *arXiv preprint arXiv:1006.3498*, 2010.

[10] P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM, 2010.

[11] Z. Guo and D. Barbosa. Robust entity linking via random walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 499–508, New York, NY, USA, 2014. ACM.

[12] X. Han and L. Sun. An entity-topic model for entity linking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 105–115. Association for Computational Linguistics, 2012.

[13] X. Han, L. Sun, and J. Zhao. Collective entity linking in web text: a graph-based method. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 765–774. ACM, 2011.

[14] Z. He, S. Liu, M. Li, M. Zhou, L. Zhang, and H. Wang. Learning entity representation for entity disambiguation. In *ACL (2)*, pages 30–34, 2013.

[15] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 782–792. Association for Computational Linguistics, 2011.

[16] N. Houlsby and M. Ciaramita. A scalable gibbs sampler for probabilistic entity linking. In *Advances in Information Retrieval*, pages 335–346. Springer, 2014.

[17] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

[18] S. S. Kataria, K. S. Kumar, R. R. Rastogi, P. Sen, and S. H. Sengamedu. Entity disambiguation with hierarchical topic models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1037–1045. ACM, 2011.

[19] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466. ACM, 2009.

[20] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[21] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM, 2007.

[22] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[23] A. Moro, A. Raganato, and R. Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244, 2014.

[24] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 467–475, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[25] F. Piccinno and P. Ferragina. From tagme to wat: a new entity annotator. In *Proceedings of the first international workshop on Entity recognition & disambiguation*, pages 55–62. ACM, 2014.

[26] A. Pilz and G. Paaß. From names to entities using thematic context distance. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 857–866. ACM, 2011.

[27] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics, 2011.

[28] B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In J. Shawe-taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 693–701. 2011.

[29] G. Rizzo, M. van Erp, and R. Troncy. Benchmarking the extraction and disambiguation of named entities on the semantic web. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, 2014.

[30] A. Sil and A. Yates. Re-ranking for joint named-entity recognition and linking. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2369–2374. ACM, 2013.

[31] S. Singh, S. Riedel, B. Martin, J. Zheng, and A. McCallum. Joint inference of entities, relations, and coreference. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 1–6. ACM, 2013.

[32] V. I. Spitkovsky and A. X. Chang. A cross-lingual dictionary for English Wikipedia concepts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May 2012.

[33] N. Steinmetz and H. Sack. Semantic multimedia information retrieval based on contextual descriptions. In *The Semantic Web: Semantics and Big Data*, pages 382–396. Springer, 2013.

[34] Y. Sun, L. Lin, D. Tang, N. Yang, Z. Ji, and X. Wang. Modeling mention, context and entity with neural networks for entity disambiguation.

[35] C. Sutton and A. McCallum. Piecewise training for structured prediction. *Machine Learning*, 77(2-3):165–194, 2009.

[36] R. Usbeck, A.-C. N. Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, and A. Both. Agdistis-graph-based disambiguation of named entities using linked data. In *The Semantic Web–ISWC 2014*, pages 457–471. Springer, 2014.

[37] R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, et al. Gerbil: General entity annotator benchmarking framework. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1133–1143. International World Wide Web Conferences Steering Committee, 2015.

[38] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy. Accelerated training of

conditional random fields with stochastic gradient methods. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 969–976, New York, NY, USA, 2006. ACM.

[39] J. Yedidia, W. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 13, pages 689–695, Dec. 2000.

[40] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.